

ORIGINAL ARTICLE

Three- and four-digit ICD-10 is not a reliable classification system in primary care

ROSEMARIE WOCKENFUSS^{1,2}, THOMAS FRESE², KRISTIN HERRMANN²,
MELANIE CLAUSNITZER² & HAGEN SANDHOLZER²

¹General Practitioner, Döbeln, ²Department of Primary Care, Leipzig Medical School, Leipzig, Germany

Abstract

Objective. The International Classification of Diseases 10th revision (ICD-10) is a standard international diagnostic classification for medical diagnoses. Reliable diagnostic coding is of high medical and epidemiological importance. Coding diagnoses with ICD-10 is the basis of reimbursement in some healthcare systems. **Design.** The ICD-10 coding of each case was performed by two raters to investigate the inter-rater agreement. The degree of agreement was assessed using Cohen's kappa. Kappa was divided into two groups: Kappa ≥ 0.61 meaning high or satisfactory and kappa ≤ 0.6 (incl. ≤ 0.000 and 0.000^*) meaning low or unsatisfactory. **Subjects.** Cross-sectional data were collected from 8877 randomly selected patients. The 209 participating general practitioners used a standardized data collection form. The first of the reasons for encounter was taken into account on new and chronic managed problems. **Results.** Kappa values were satisfactory on the chapter level with on average 0.685 (chronic managed problems) and 0.675 (new managed problems). Kappa was unsatisfactory when the three-digit level was used (0.428) and lower for terminal codes (four-digit level) at 0.199 on average (chronic managed problems). For new managed problems the kappa values were at 0.384 (three-digit level) and 0.166 (four-digit level) on average. **Conclusion.** The ICD-10 is reliable for coding managed problems on the chapter level. Further refinement of ICD-10 with three- and four-digit codes leads to significant coding uncertainties. There is no reliable coding scheme that meets the demands of general practice. The use of coded data for healthcare reimbursement requires a simplification of ICD-10 to provide a realistic picture of morbidity.

Key Words: Family practice, general practitioner, ICD-10, primary care, reliability

The International Classification of Diseases and Health Problems 10th revision (ICD-10) is a standard for classifying diagnoses, symptoms, and other medical care encounters [1]. The International Classification of Primary Care (ICPC) is also a widespread classification system [2]. Reliable encoding of diagnoses is of high medical and epidemiological importance. The increasing importance of the imaging of patients' morbidity is shown by the fact that reimbursement in Germany's ambulatory healthcare system became morbidity-based. Besides age and gender, ICD-10 coded diagnoses will be the basis for this reimbursement and thus their exact coding will be of existential relevance. There are three volumes of the ICD-10: Volume 1: Tabular list, volume 2: Instruction manual, volume 3: Alphabetical index. The 21 chapters are subdivided into

homogenous "blocks" of three alphanumeric-character categories. This is the "core" classification of ICD-10. The four-character subcategories are recommended for many purposes and form an integral part of the ICD, as do the special tabulation lists. There have only been a few studies investigating the reliability of ICD-10 in the primary healthcare context.

Material and methods

The Saxon Society of General Medicine (SGAM) contacted all general practitioners in Saxony. Some 270 declared their willingness to participate and 209 of the 2510 physicians cooperated. Cross-sectional data were collected from 1 October 1999 to 30 September 2000. Case recording was carried out on

Correspondence: Dr H. Sandholzer, Department of Primary Care of the Leipzig Medical School, Philip-Rosenthal Straße 27a, 04103 Leipzig, Germany.
E-mail: sanh@medizin.uni-leipzig.de

(Received 12 December 2008; accepted 8 May 2009)

ISSN 0281-3432 print/ISSN 1502-7724 online © 2009 Informa UK Ltd. (Informa Healthcare, Taylor & Francis AS)
DOI: 10.1080/02813430903072215

The reliability of the ICD-10 in primary care has not yet been investigated. Extensive investigations were performed on cross-sectional data from 8877 primary care patients.

- Three- and four-digit ICD-10 is not reliable in primary care.
- ICD-10 is reliable only at chapter level.
- Small and easy terminologies are more appropriate for primary care.
- Clear coding rules should be established.
- Further coding refinement increases the error rate.

one day a week (Monday to Friday; either morning or afternoon consultation hours), chosen at random. Data were collected for one in 10 patients previously known to the practitioner. Multiple screenings of the same patient were avoided. House calls were not considered. A total of 8877 patients were included. A standardized data collection form was used. It was developed by general practitioners (Leipzig Medical School and Saxon Society of General Medicine). The form was tested and evaluated during a pilot trial (SESAM 1). It was found to be relevant, and cost- and time-efficient. Each patient's reasons for the encounter, symptoms, diagnostic procedures, recent diagnoses, and general morbidity were estimated as well as therapeutic procedures. As far as possible, data were literally documented (according to the study instructions), either as stated by the patients (e.g. reasons for encounter) or in the words of the physician (e.g. chronic diagnoses). Due to the randomization pattern, the information was documented with a passable expenditure of time. Only fully completed forms were considered. Data were categorized using the "International Classification of Diseases" (10th revision, ICD-10). These data were edited by two groups of medical doctoral candidates (both sexes, specialized general practitioners with their own medical practice, educated at different universities, age approximately 30 to 50 years) of Leipzig Medical School's Department of General Practice. To evaluate the reliability of ICD-10 as a classification instrument the same data were always been coded by two raters in parallel. Only the first reason for the encounter, i.e. the first chronic managed problem or the first new managed problem, was considered. Following the classification by Landis and Koch [3] kappa value agreement rating has been slightly simplified by bisection: Kappa \geq 0.61 was to be considered high or satisfactory; kappa \leq 0.6 was to be considered low or

Table I. Averaged number of cases in each ICD-10 chapter.

ICD-10 chapter		n (chronic)	n (new)
I	Infectious diseases	41	161.5
II	Neoplasms	202.5	39.5
III	Blood diseases	44.5	33.5
IV	Endocrine & metabolism	1168.5	216
V	Mental & behavioural	319.5	205.5
VI	Nervous system	179	129
VII	Eye diseases	23	47.5
VIII	Ear diseases	36	94.5
IX	Circulatory system	2817	533.5
X	Respiratory system	454	1235
XI	Digestive system	331.5	433
XII	Skin diseases	177.5	214.5
XIII	Musculoskeletal system	829	1130
XIV	Genitourinary system	92.5	146.5
XV	Pregnancy	7.5	4
XVII	Congenital malformations	34.5	5
XVIII	Symptoms & findings	68.5	222
XIX	Injuries	87.5	423
XX	External causes	0.5	351.5
XXI	Health status	17.5	161.5

unsatisfactory. No real kappa value could be calculated for codes assigned by only one of the raters. In these cases a hypothetical value of kappa = 0 was assumed, marked as 0.000* and declared unsatisfactory.

Results

Data on 8877 patients were estimated. The average number of cases belonging to each ICD-10 chapter is given in Table I. In terms of coding reliability with regard to chronic managed problems, a high degree of agreement with kappa > 0.6 was ascertained in 14 (65%) of 20 ICD-10 chapters (average kappa value of all chapter kappa values was 0.685), with three-digit coding in two (10%) chapters (average kappa value of all three-digit codes used was 0.428; a high degree of agreement in 42.93% of all three-digit codes used) and with four-digit coding in no chapter (average kappa value of all four-digit codes used 0.199; high degree of agreement in 18.02% of all four-digit codes used). In chapter XX, "External causes", no kappa value was determined as only one single code was assigned by a rater. Unsatisfactory agreement was found in three chapters. In no chapter was a high degree of agreement determined for four-digit coding. These results are summarized in Figure 1.

A high agreement (kappa > 0.61) was detected in 12 of 19 chapters (63.16%) for new managed problems coding (average kappa value of all chapter kappa: 0.675). In no chapter was agreement satisfactory for three- or four-digit coding (average kappa

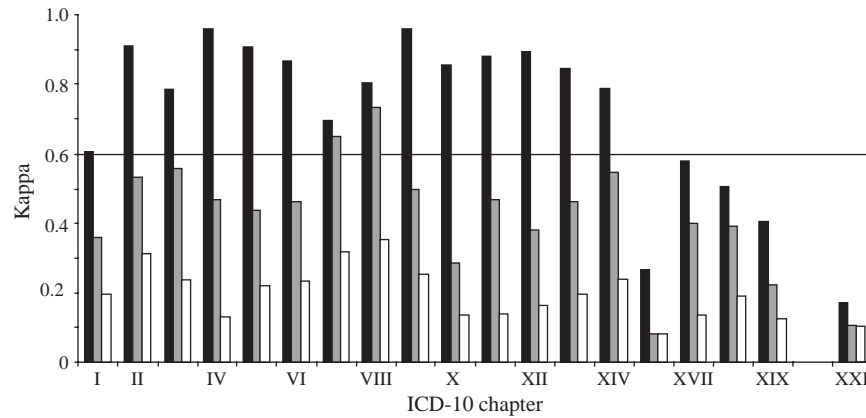


Figure 1. Chapter kappa (black) and average kappa value of all three-digit (grey) and four-digit (transparent) codes used chronic managed problems (chapters XV and XVI were summarized).

value of all three-digit codes used 0.384; high degree of agreement in 36.01% of all three-digit codes used and average kappa value of all four-digit codes used 0.166 with a high degree of agreement in 11.85% of all four-digit codes used). Satisfactory degrees of agreement with three-digit coding were found in none of the 19 chapters; the highest agreement was noted in chapters VIII “Ear diseases” (kappa mean value 0.553), VII “Eye diseases” (kappa mean value 0.543), XIV “Genitourinary system” (kappa mean value 0.491) and III “Blood diseases” (kappa mean value 0.479). These values are based on a small amount of data (see Table I) that inadequately represent patients in general practice. In chapters XXI “Factors influencing health status and contact with health services” (kappa: 0.175) and IV “Endocrine, nutritional and metabolic diseases” (kappa 0.295) a very small consensus between the raters was ascertained. With four-digit coding no satisfactory agreement could be found. These results are summarized in Figure 2.

Discussion

The SESAM 2 study was conducted independently of industrial sponsorship; the results can be assumed to be representative and uninfluenced by attention, response, or seasonal bias [4]. The patients were randomly selected. A total estimation of all patients could consider the order of patients within the consultation hour. It was impossible for the raters to be present at all the 8877 consultations. Video-taping of the consultations was also impractical. Electronic patient records are not a good documentation method [5]. This is why the described randomization and documentation were performed.

In preparation of the proposals regarding the International Classification of Diseases and Health Problems 10th revision (ICD-10) particular importance was placed on the structural examination of ICD-10 with regard to the matter of the fundamental appropriateness of ICD-10 for disease classification and medical conditions and thus to what

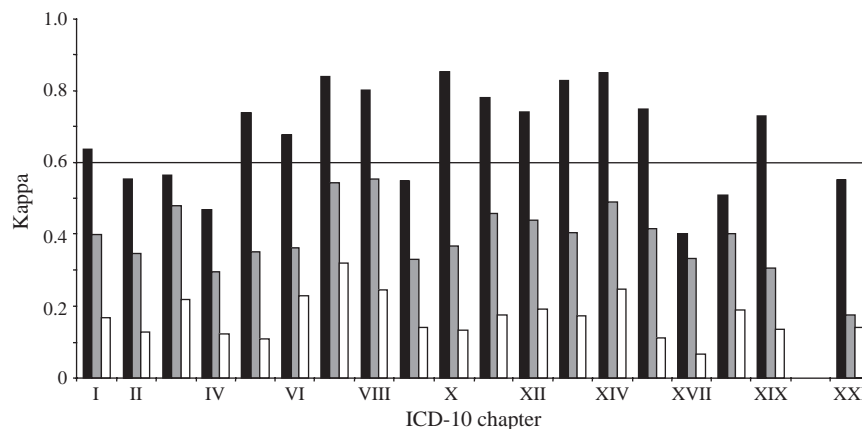


Figure 2. Chapter kappa (black) and average kappa value of all three-digit (grey) and four-digit (transparent) codes used new managed problems (chapters XV and XVI were summarized).

extent it meets the manifold demands imposed on morbidity and healthcare statistics. Bernstein et al. stated [6], “The most important aspect of reliability testing is agreement in classification” and drew the conclusion that in primary healthcare small simple terminologies are particularly suitable. Surjan [7] argues that the error rate rises with higher-level coding refinement. So do Stausberg et al. [8]. They consider the refinement of ICD-10 as problematic, even leading to coding uncertainties among experts. There is a demand for simplification of ICD-10. We show that coding agreement is much better on the chapter level than with three-digit coding, which again is better than four-digit coding. It should be taken into consideration that kappa values do not represent the quality of the diagnoses – a high kappa value could result from two wrong but matching diagnoses. However, a fundamental coding error by two independent raters seems to be an unlikely event. The reliability of coding would be underestimated by a coding failure from one of the raters.

The SESAM 2 study shows lower kappa values than other studies. This might be caused by the fact there is a broad spectrum of diagnoses. Research on ICD-10 reliability has so far focused on inpatient healthcare or chapter V “Mental diseases”. Stausberg et al. [8,9] found kappa values between 0.27 (fair) and 0.42 (moderate) for reliability comparison with four-digit coding. Coding on the chapter level achieved substantial results, with kappa values between 0.71 and 0.72. Four-digit coding of all diagnoses showed a lower degree of agreement (kappa 0.21) than coding of only one main diagnosis (kappa 0.29). Stausberg and Lehmann [10] entrusted students with coding to draw conclusions about amateur coding quality. The reliability among the raters, who used a newly introduced, simplified four-digit coding benchmark, amounted to 0.46 compared with 0.87 at chapter level.

In attempting to combine the categories of the “Primary Health Care Version of the International Statistical Classification of Diseases and Related Health Problems” (KSH97-P PHC) and “Systematized Nomenclature of Medicine–Clinical Terms” (SNOMED–CT) terms, Vikstrom et al. [11] found that consideration of clear coding rules achieved good inter-coder reliability (improvement from 69% to 83%). Examining mental diseases’ coding reliability for 18-month-old children Skovgaard et al. [12] showed a kappa value on axis I of 0.66 by using ICD-10 and 0.72 coding with DC 0-3. An international field study to verify the multiaxial system of ICD-10’s chapter F compared the reliability of the three axes [13]. Diagnoses on axis I (psychiatric and somatic diagnoses) showed a kappa value of 0.50; on axis II (psychosocial function impairment) the

intra-class coefficient measuring the reliability among the raters was calculated to be 0.62. Both results evidence medium reliability. Kappa value on axis III (load factors) of 0.16 reached only low reliability. Nilsson et al. compared the reliability of KSH97-P, which is a short version of ICD-10 for general healthcare, using a trisection into book version, computer version with traditional ICD-10 structure, and computer version with composed ICD-10 structure [14]. Reliability was poor on code level (kappa between 0.53 and 0.58) in all three versions, but on chapter level kappa values between 0.76 and 0.82 could be reached. In the recent investigation kappa values were lower and ranged from 0.675 (new managed problems) to 0.685 (chronic managed problems) on chapter level. Goldstein et al. identified coding error rates of 29% to 50% at a kappa value of 0.68 [15]. Except for bipolar dysfunctions having been the other way round, Hiller et al. found a generally higher reliability of ICD-10 (kappa 0.59) than of the Diagnostic and Statistical Manual of Mental Disorders (DSM)-III-R (kappa 0.53; [16]).

Using the example of eating disorders, Nicholls et al. examined the reliability of three different classification systems and found the best analogy (kappa 0.879) using the Great Ormond Street (GOS) criteria, specially made for that purpose [17]. DSM-IV classification showed a kappa rate of 0.636 and ICD-10 a kappa of 0.357. Willemse et al. checked the coding reliability on ICD-10’s psychosocial axis and detected only moderate reliability (kappa < 0.61) in most cases [18]. Gibson and Bridgman found an error rate of 29% [19], 8% of the errors were derived from choosing the wrong chapter and 15% from coding with three digits. Henderson et al. evaluated ICD-10-AM and its coding quality [20]. Weighted kappa values of around 0.9 resulted. Thus kappa values have been higher than in all other studies [21–23].

Our investigations showed a moderate reliability and a relevant number of coding errors when coding with ICD-10. The recent data were estimated in a daily setting. When reimbursement becomes morbidity based, a reliable coding system is existential. This is a fundamental circumstance for all general practitioners. Further investigations on the International Classification System of Primary Care have been performed with the recent data. The results will show the significance of the ICPC as an alternative coding system.

Strengths of the study

- The study investigates a broadly used classification system (ICD-10).

- The study investigates the reliability of ICD-10 in a primary care setting.
- The study includes all groups of diagnoses.

Weaknesses of the study

- The raters could not be present at the consultations.
- Only about 10% of the general practitioners cooperated.
- Data could not be estimated from all patients.

Conclusion

We determined ICD-10 reliability in primary health-care. We have shown that mean reliability of chronic managed problems and new managed problems is still satisfactory on the chapter level whereas three- or four-digit coding provides unsatisfactory results. The conclusion has to be that small and easy terminologies are more appropriate for primary healthcare as further coding refinement on a higher level increases the error rate. Thus ICD-10 simplification and clear coding rules are required.

Competing interests/funding of the studies/participation

The authors declare that they have no competing interests. Their investigations were funded by the Leipzig Medical School and the Saxon Society of General Medicine.

RW wrote the initial manuscript, TF participated in the data collection and revision of the manuscript, KH performed the statistical analysis, MC and RW converted the free text into ICD-10 codes, HS participated in the design of the study and reviewed the manuscript.

Acknowledgements

The authors would like to thank Dr Hanno Grethe, honorary president of the SGAM, and Dr Johannes Dietrich, president of the SGAM, for their kind support.

References

- [1] Koch H, Graubner B, Brenner G. Erprobung der Diagnosenverschlüsselung mit der ICD-10 in der Praxis des niedergelassenen Arztes. ZI für die kassenärztliche Versorgung in der BRD. Wissenschaftliche Reihe Band 54. Köln: Deutscher Ärzte-Verlag; 1998.
- [2] Soler JK, Okkes I, Wood M, Lamberts H. The coming of age of ICP: Celebrating the 21st birthday of the International Classification of Primary Care. *Fam Pract* 2008;25:312–7.
- [3] Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
- [4] Frese T, Sandholzer H, Voigt S, Voigt R. Epidemiology of diabetes mellitus in German general [r]actioners' consultation – Results of the SESAM 2 study. *Exp Clin Endocrinol Diabetes* 2008;116:326–8.
- [5] Vainiomäki S, Kuusela M, Vainiomäki P, Rautava P. The quality of electronic patient records in Finish primary healthcare needs to be improved. *Scand J Prim Health Care* 2008;26:117–22.
- [6] Bernstein RM, Hollingworth GR, Viner G, Shearman J, Labelle C, Thomas R. Reliability issues in coding encounters in primary care using an ICP/ICD-10-based controlled clinical terminology. *J Am Med Informatics Assoc Symposium (Suppl)* 1997;21:843.
- [7] Surján G. Questions on validity of International Classification of Diseases-coded diagnoses. *Int J Med Inform* 1999;54:77–95.
- [8] Stausberg J, Lehmann N, Kaczmarek D, Stein M. Reliability of diagnoses coding with ICD-10. *Int J Med Inform* 2008;77:50–7.
- [9] Stausberg J, Lehmann N, Kaczmarek D, Stein M. Einheitsliches Kodieren in Deutschland: Wunsch und Wirklichkeit. *Das Krankenhaus* 2005;8:657–62.
- [10] Stausberg J, Lehmann N. Kodierübungen im Medizinstudium. Wie gut kodieren Anfänger mit der ICD-10? *GMS Med Inform Biom Epidemiol* 2005;1:Doc04
- [11] Vikström A, Skånér Y, Strender LE, Nilsson GH. Mapping the categories of the Swedish primary health care version of ICD-10 to SNOMED CT concepts: Rule development and intercoder reliability in a mapping trial. *BMC Med Inform Decis Mak* 2007;2:7–9.
- [12] Skovgaard AM, Houmann T, Christiansen E, Andreasen AH. The reliability of the ICD-10 and the DC 0-3 in an epidemiological sample of children 1½ years of age. *Infant Mental Health J* 2005;26:470–80.
- [13] Siebel U, Michels R, Hoff P, Schaub RT, Droste R, Freyberger HJ, Dilling H. Multiaxiales System des Kapitels V(F) der ICD-10 Erste Ergebnisse der der multizentrischen Praktikabilitäts- und Reliabilitätsstudie. *Nervenarzt* 1997;68:231–8.
- [14] Nilsson G, Petersson H, Ahlfeldt H, Strender LE. Evaluation of three Swedish ICD-10 primary care versions: Reliability and ease of use in diagnostic coding. *Methods Inf Med* 2000;39:325–31
- [15] Goldstein LB, Jones MR, Matchar DB, Edwards LJ, Hoff J, Chilukuri V, Armstrong SB, Horner RD. Improving the reliability of stroke subgroup classification using the Trial of ORG 10172 in Acute Stroke Treatment (TOAST) criteria. *Stroke* 2001;32:1091–8
- [16] Hiller W, Dichtl G, Hecht H, Hundt W, von Zerssen D. An empirical comparison of diagnoses and reliabilities in ICD-10 and DSM-III-R. *Eur Arch Psychiatry Clin Neurosci* 1993;242:209–17
- [17] Nicholls, Chater R, Lask B. Children into DSM don't go: A comparison of classification systems for eating disorders in childhood and early adolescence. *Int J Eat Disord* 2000;28:317–24.
- [18] Willemse GR, van Yperen TA, Rispens J. Reliability of the ICD-10 classification of adverse familial and environmental factors. *J Child Psychol Psychiatry* 2003;44:202–13
- [19] Gibson N, Bridgman SA. A novel method for the assessment of the accuracy of diagnostic codes in general surgery. *Ann R Coll Surg Engl* 1998;80:293–6.

- [20] Henderson T, Shepherd J, Sundararajan V. Quality of diagnosis and procedure coding in ICD-10 administrative data. *Med Care* 2006;44:1011–9.
- [21] Humphries KH, Rankin JM, Carere RG, Buller CE, Kiely FM, Spinelli JJ. Co-morbidity data in outcomes research: Are clinical data derived from administrative databases a reliable alternative to chart review? *J Clin Epidemiol* 2000; 53:343–9.
- [22] Quan H, Parsons GA, Ghali WA. Validity of information on comorbidity derived from ICD-9-CCM administrative data. *J Clin Epidemiology Med Care* 2002;40:675–85.
- [23] Quan H, Parsons GA, Ghali WA. Validity of procedure codes in International Classification of Diseases, 9th revision, Clinical modification administrative data. *Med Care* 2004;42:801–9.